Analysis of Codon Usage Bias in the Chloroplast Genome of Eucalyptus grandis (Postprint)

Authors: Wang Pengliang, Wu Shuangcheng, Yang Liping, Wang Huayu, Chen Naiming, Zhang Zhaoyuan

Date: 2019-01-30T00:00:00+00:00

Abstract

To improve gene expression efficiency and thereby utilize chloroplast genetic engineering to modify important traits of Eucalyptus grandis, this study employed the chloroplast genome sequence of Eucalyptus grandis as material, selecting 43 non-redundant genes longer than 300 nt with AUG as the start codon as research subjects, and analyzed codon usage bias in the Eucalyptus grandis chloroplast genome using CodonW 1.4.2 software. The analysis revealed: the average GC content at the third codon position was 27.97%; ENC ranged from 39.49 to 61.00, averaging 47.04; there were 31 codons with RSCU > 1, of which 29 ended with A/U; neutrality analysis showed no significant correlation between GC12 and GC3, and regression analysis also failed to reach significance; ENC-plot analysis revealed that most genes fell on or near the curve; correspondence analysis indicated that the first axis contributed 17.68%, the second axis 11.49%, the third and fourth axes 8.00% and 5.76%, respectively, with a cumulative contribution of 42.93% for the first four axes, and the first axis was extremely significantly correlated with GC, ENC, and CAI. These results demonstrate that codon bias in the Eucalyptus grandis chloroplast genome is weak, with the third codon position preferentially ending with A or U; selection and mutation play relatively balanced roles in shaping codon bias in the Eucalyptus grandis chloroplast genome; ultimately, 12 high-frequency, high-expression codons—UUG, CUU, GUU, UCC, UCA, ACA, UAU, UAA, CAU, AAU, AGA, and GGA—were identified as optimal codons. This study lays a solid foundation for optimizing chloroplast gene codons to enhance expression efficiency and modify target traits in Eucalyptus grandis.

Full Text

Analysis of Codon Bias in the Chloroplast Genome of *Eucalyptus grandis*

Pengliang Wang[1,2], Shuangcheng Wu[2], Liping Yang[3], Huayu Wang[2], Naiming Chen[3], Zhaoyuan Zhang[1*]

[1]Guangxi Key Laboratory of Superior Timber Trees Resource Cultivation, Guangxi Forestry Research Institute, Nanning 530002, China
[2]Guangxi Key Laboratory of Beibu Gulf Marine Biodiversity Conservation, Qinzhou University, Qinzhou 535011, Guangxi, China
[3]Qinzhou Key Laboratory of Plant Biotechnology, Qinzhou Forestry Research Institute, Qinzhou 535099, Guangxi, China

Abstract

To enhance gene expression efficiency for chloroplast genetic engineering aimed at improving important traits in *Eucalyptus grandis*, this study analyzed codon usage bias in the chloroplast genome of *E. grandis*. Using the chloroplast genome sequence as source material, we selected 43 non-redundant genes longer than 300 nt that initiate with AUG as study objects and employed CodonW 1.4.2 software to analyze codon usage preferences. The results revealed that the average GC content at the third codon position was 27.97%. The effective number of codons (ENC) ranged from 39.49 to 61.00, with a mean of 47.04. Among the 31 codons with RSCU values greater than 1, 29 ended with A or U. Neutrality plot analysis showed no significant correlation between GC12 and GC3, and regression analysis also failed to reach significance. ENC-plot analysis demonstrated that most genes fell on or near the standard curve. Correspondence analysis indicated that the first axis contributed 17.68% of the variation, the second axis 11.49%, and the third and fourth axes 8.00% and 5.76%, respectively, with the first four axes cumulatively accounting for 42.93% of the total variation. The first axis showed highly significant correlations with GC, ENC, and CAI. These findings indicate that codon bias in the *E. grandis* chloroplast genome is relatively weak, with a preference for A or U at the third codon position. Both selection and mutation appear to play relatively balanced roles in shaping codon preferences. Ultimately, twelve high-frequency, high-expression codons were identified as optimal: UUG, CUU, GUU, UCC, UCA, ACA, UAU, UAA, CAU, AAU, AGA, and GGA. This research establishes a solid foundation for optimizing codons in transgenes to improve expression efficiency and enhance target traits through chloroplast genetic engineering in *E. grandis*.

Keywords: *Eucalyptus grandis*, chloroplast, genome, codon bias

Introduction

Eucalyptus grandis, a perennial woody species in the Myrtaceae family, is native to Australia. Due to its rapid growth, straight trunk form, and tall stature, it has been introduced and widely cultivated worldwide, becoming an important exotic species in many countries [1]. Building upon introduction and domestication efforts, researchers have conducted extensive studies on variation in different traits among provenances, families, and individual trees. These investigations have revealed that *E. grandis* exhibits insufficient cold resistance [12], is susceptible to gall wasp (*Leptocybe invasa*) infection [24], and shows substantial variation in growth and stem form characteristics among different genetic resources [19, 23].

Genetic engineering offers significant advantages over traditional breeding, including stronger targeting, shorter cycles, and higher efficiency [15]. Chloroplast genetic engineering, in particular, enables high-level transgene expression while effectively controlling transgene dispersal, making it an ideal transformation approach [2]. Codon usage, often termed the "second genetic code" [13, 3], influences not only gene expression levels [26] but also corresponding protein functions [4]. Substantial differences in chloroplast genome codon preferences exist among different species [25, 11, 16, 17]. This study aims to characterize codon usage bias in the *E. grandis* chloroplast genome and identify its optimal codons, thereby laying the groundwork for chloroplast genetic engineering and genetic improvement of this important timber species.

---

1. Materials and Methods

1.1 Sequence Data The chloroplast genome of *Eucalyptus grandis* was retrieved from the NCBI organelle genome database (https://www.ncbi.nlm.nih.gov/nuccore/NC_014570.1) by searching for the Latin name *Eucalyptus grandis*. The complete genome sequence in FASTA format and coding sequences (CDS) were downloaded. The *E. grandis* chloroplast genome spans 160,137 bp and contains 75 genes. To minimize analytical errors, we selected 43 non-redundant sequences longer than 300 nt that initiate with AUG for codon bias analysis.

1.2 Data Analysis

1.2.1 Calculation of Codon Bias Parameters The coding sequences of the 43 selected non-redundant genes were analyzed using CodonW 1.4.2 software to calculate codon usage parameters: relative synonymous codon usage (RSCU), effective number of codons (ENC) with a theoretical minimum of 20 (indicating only one codon per amino acid) and maximum of 61 (indicating equal usage of all codons), codon adaptation index (CAI, ranging from 0 to 1 with higher values indicating stronger bias), codon bias index (CBI), frequency of optimal codons (FOP), protein hydrophobicity (Gravy), and GC content at different positions

including GC1, GC2, GC3, GC3S, GC12, and overall GC content (representing GC content at the first, second, and third codon positions, synonymous third position, average of first and second positions, and total codon GC content, respectively).

1.2.2 Neutrality Plot Analysis To preliminarily identify factors influencing codon bias, neutrality plot analysis was performed by calculating the average of GC1 and GC2 (GC12) for the y-axis and plotting against GC3 on the x-axis. Genes positioned on the diagonal line indicate mutational effects, while those deviating from the diagonal suggest selection pressure, thereby revealing the primary forces shaping codon usage preferences.

1.2.3 ENC-plot Analysis To further elucidate codon bias determinants, ENC-plot analysis was conducted with ENC on the y-axis and GC3S on the x-axis. Genes were plotted as scatter points, and the standard curve was added according to Wright's equation [18]:

$$ENC = 2 + GC3S + \frac{29}{GC3S^2 + (1 - GC3S)^2}$$

The ENC ratio was calculated as:

$$\text{ENC ratio} = \frac{ENC_{\text{observed}} - ENC_{\text{expected}}}{ENC_{\text{expected}}}$$

Based on the distribution of scatter points and ENC ratios, deviation from the standard curve indicates selection pressure, while alignment with the curve suggests mutational effects, allowing inference of the likely causes of codon bias.

1.2.4 Correspondence Analysis Correspondence analysis employs appropriate scaling methods to combine variable and sample analysis, simultaneously revealing relationships between samples and variables on the same factor plane. We applied this method using CodonW software to analyze codon usage patterns in the *E. grandis* chloroplast genome and uncover underlying usage principles.

1.2.5 Determination of Optimal Codons To identify optimal codons, all test genes were ranked by ENC values. The top and bottom 10% of genes were selected to construct high-expression and low-expression gene libraries, respectively. Codons with ΔRSCU (difference in RSCU between high- and low-expression libraries) greater than 0.08 and RSCU values exceeding 1 were designated as optimal codons [9, 21, 22, 16].

---

2. Results and Analysis

2.1 Codon Composition Analysis To accurately analyze codon bias, we selected 43 non-redundant genes from the *E. grandis* chloroplast genome that initiate with AUG and have coding sequences longer than 300 nt. CodonW software was used to calculate relevant parameters for these genes. The results (Table 1) revealed variation in GC content across codon positions: GC1 ranged from 34.20% to 58.90% (mean 47.40%), GC2 from 27.90% to 58.70% (mean 39.47%), and GC3 from 20.20% to 37.00% (mean 27.97%). Notably, GC content at the first and second positions was substantially higher than at the third position. ENC values ranged from 39.49 to 61.00 with an average of 47.04. CAI varied between 0.082 and 0.301 (mean 0.1714), CBI from -0.222 to 0.196 (mean -0.092), and FOP from 0.263 to 0.532 (mean 0.356). Protein hydrophobicity (Gravy) ranged from -0.704 to 1.102 with a mean of 0.017.

Correlation analysis of codon parameters (Table 2) showed significant positive correlation between GC1 and GC2 ($r = 0.363$). However, neither GC1 nor GC2 correlated significantly with GC3. Overall GC content correlated highly significantly with GC1 and GC2 but not with GC3. ENC showed no correlation with GC1, significant negative correlation with GC2, and highly significant positive correlation with GC3 ($r = 0.521$). Both GC1 and overall GC correlated highly significantly with CAI, CBI, and FOP, while GC3 showed significant correlation with these indices. Gravy did not correlate significantly with any other codon parameters. Codon number (N) correlated highly significantly with GC3 but not with ENC, CAI, or other parameters.

RSCU analysis (Table 3) identified 31 codons with RSCU values exceeding 1.00. Among these, 16 codons ended with U, 13 ended with A, and only one each ended with G or C. Codons ending with A or U accounted for 93.54% of all biased codons.

2.2 Neutrality Plot Analysis Neutrality plot analysis of *E. grandis* chloroplast genes revealed that GC12 ranged from 33.65% to 55.45%, while GC3 ranged from 20.20% to 37.00%. The correlation between GC12 and GC3 was not statistically significant, indicating weak association between these parameters and differential mutational effects on base composition across codon positions. If codon usage were determined solely by random mutation, genes would distribute along the diagonal line. However, Figure 1 shows that most genes clustered above the diagonal, with GC12 consistently higher than GC3, suggesting that selection plays a predominant role in shaping codon bias.

2.3 ENC-plot Analysis ENC-plot analysis positioned all test genes in a coordinate system with ENC on the y-axis and GC3S on the x-axis, incorporating the standard curve according to equation (1). The results (Figure 2) showed that while a small subset of genes deviated from the standard curve, most located near it. To quantify these deviations, we calculated theoretical ENC values using equation (1) and derived ENC ratios using equation (2). Frequency dis-

tribution analysis of ENC ratios (Table 4) revealed that 51.16% of genes fell within the -0.05 to 0.05 range, 34.88% within 0.05 to 0.15, 9.30% within -0.15 to -0.05, and the remaining 2.33% within -0.25 to -0.15 or 0.15 to 0.25. These results indicate that mutation plays an important role in shaping codon bias in the *E. grandis* chloroplast genome.

2.4 Correspondence Analysis Correspondence analysis revealed that the first axis accounted for 17.68% of the variation, the second axis 11.49%, and the third and fourth axes 8.00% and 5.76%, respectively. The cumulative contribution of the first four axes reached 42.93%. Since both the first and second axes contributed over 10% each, they represent major factors influencing codon bias. The first axis correlated highly significantly and positively with GC, CAI, CBI, and Fop ($r = 0.573, 0.670, 0.578$, and 0.523, respectively) and highly significantly negatively with ENC ($r = -0.395$). Although the first axis did not correlate significantly with GC3S, it showed highly significant correlations with A and G content at synonymous third positions ($r = -0.440$ and -0.606, respectively). To visualize codon preferences, we constructed a planar coordinate system with the first axis as x and the second axis as y, plotting genes according to their functional categories (Figure 3). Ribosomal protein genes clustered relatively tightly, while other genes distributed more dispersedly, indicating similar codon preferences among ribosomal protein genes but distinct preferences compared to other gene categories.

2.5 Determination of Optimal Codons Using ENC values as the criterion, we ranked all test genes and selected the top and bottom 10% (four genes each) to construct high- and low-expression gene libraries. RSCU values were recalculated for each library, and ΔRSCU values were derived (Table 5). Applying the $\Delta\text{RSCU} > 0.08$ criterion identified 31 high-expression codons (marked with * in Table 5), including 12 ending with G, 8 with C, 6 with A, and 5 with U.

By comparing high-frequency codons from Table 3 with high-expression codons from Table 5, we identified shared codons as optimal. Twelve optimal codons were determined for the *E. grandis* chloroplast genome: UUG, CUU, GUU, UCC, UCA, ACA, UAU, UAA, CAU, AAU, AGA, and GGA. Ten of these end with U or A, while two end with G or C.

Discussion

The genetic code defines the correspondence between nucleotide sequences and amino acid sequences. Among the 20 proteinogenic amino acids, methionine (Met) and tryptophan (Trp) are each encoded by a single codon, while the remaining 18 amino acids are encoded by 2–6 synonymous codons each—a property known as codon degeneracy [27]. Synonymous codons primarily differ at the third position. In this study, GC3 showed no significant correlation with GC1 or GC2 in the *E. grandis* chloroplast genome and was substantially lower

than both, indicating a preference for A- and U-ending codons. RSCU analysis quantitatively confirmed this pattern, consistent with reported features of chloroplast genomes in *Scutellaria baicalensis* [17], *Camellia oleifera* [16], and *Medicago truncatula* [22].

Codon usage bias refers to the phenomenon where organisms preferentially use specific synonymous codons during protein synthesis [20]. The average ENC value of 47.04 for the *E. grandis* chloroplast genome exceeds the threshold of 35, where values below 35 indicate strong bias and values above 35 indicate weak bias [8]. Therefore, the *E. grandis* chloroplast genome exhibits weak codon bias, a conclusion supported by CAI analysis.

Codon bias is influenced by multiple factors including base composition, selection pressure, tRNA abundance, gene length, and protein hydrophobicity [10]. In this study, codon number did not correlate significantly with ENC or CAI, suggesting gene length has minimal effect on codon bias in the *E. grandis* chloroplast genome. Similarly, protein hydrophobicity showed no significant influence. However, the highly significant correlations between the first axis and A/G content at synonymous third positions and overall GC content indicate that base composition differences affect codon preferences. The highly significant correlations of CAI, CBI, and FOP with the first axis suggest that genes selectively use codons corresponding to abundant tRNAs, leading to high expression levels [5, 6, 7], confirming that selection is an important factor. Since the correlation coefficients of A/G content and GC content with the first axis are comparable to those of CAI, CBI, and FOP, mutation and natural selection appear to play roughly equivalent roles. While neutrality plot analysis suggests selection is the relatively dominant factor, ENC-plot analysis indicates mutation also contributes substantially. We therefore conclude that mutation and selection likely play relatively balanced roles in shaping codon bias in the *E. grandis* chloroplast genome.

This study identified twelve optimal codons based on high frequency and high expression: UUG, CUU, GUU, UCC, UCA, ACA, UAU, UAA, CAU, AAU, AGA, and GGA. The determination of these optimal codons provides a solid foundation for optimizing transgene codons to enhance expression efficiency and improve important traits through chloroplast genetic engineering in *Eucalyptus grandis*.

References

References

[1] Chen SX, Zheng JQ, Liu XF, et al., 2018. Hundred year histories and prospect of Eucalyptus cultivation technology development in China. World Forestry Research, 31:7-21.

[2] Daniell H, Chase C, 2004. Molecular biology and biotechnology of plant organelles. Dordrecht: Springer.

[3] Hanson G, Coller J, 2018. Codon optimality, bias and usage in translation and mRNA decay. Nat Rev Mol Cell Biol, 19:20-30.

[4] Hershberg R, Petrov DA, 2008. Selection on codon bias. Annu Rev Genet, 42:287-299.

[5] Ikemura T, 1981a. Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes. J Mol Biol, 146:1-21.

[6] Ikemura T, 1981b. Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codon in its protein genes: a proposal for a synonymous codon choice that is optimal for the E. coli translation system. J Mol Biol, 151:389-409.

[7] Ikemura T, 1985. Codon usage and tRNA content in unicellular and multicellular organisms. Mol Biol Evol, 2:13-34.

[8] Jiang Y, Deng F, Wang H, et al., 2008. An extensive analysis on the global codon usage pattern of baculoviruses. Arch Virol, 153:2273-2282.

[9] Li J, Xue QZ, 2005. Comparison of MADS transcriptional factor on codon bias in arabidopsis and rice. J Zhejiang Univ (Agric Life Sci Ed): 513-517.

[10] Liang FF, 2010. Influencing of codon bias and its research significance. Anim Husb Feed Sci, 31(1):118-119.

[11] Liu H, Wang MX, Yue WJ, 2017. Analysis of codon usage in the chloroplast genome of Broomcorn millet (*Panicum miliaceum* L.). Plant Sci J, 35:362-371.

[12] Liu J, Xiang DY, Chen JB, et al., 2009. Low temperature LT50 of three Eucalyptus seedlings with Electrical conductivity method and logistic equation. Guangxi Forestry Science, 38:75-78.

[13] Nelson DL, Cox MM, 2017. Lehninger Principles of Biochemistry. New York: W.H. Freeman and Company.

[14] Qi SX, 2006. Introduction and status of Eucalyptus in China. Guangxi Forestry Science, 35:250-252.

[15] Wang GL, Fang HJ, 2014. Plant genetic engineering. Beijing: Science Press.

[16] Wang PL, Yang LP, Wu HY, et al., 2018. Codon preference of chloroplast genome in *Camellia oleifera*. Guihaia, 38:135-144.

[17] Wang WB, Yu H, Qiu XP, 2018. Analysis of repeat sequence and codon bias of chloroplast genome in *Scutellaria baicalensis*. Molecular Plant Breeding, 16:2445-2452.

[18] Wright F, 1990. The effective number of codons used in a gene. Gene, 87:23-29.

[19] Wu SJ, Chen GC, Xu JM, et al., 2016. Variation analysis and selection for *Eucalyptus grandis* provenances and families in multiple-sites. For Environ Sci, 32:10-15.

[20] Wu XM, Wu SF, Ren DM, et al., 2007. The analysis method and progress in the study of codon bias. HEREDITAS, 29:420-426.

[21] Xu C, Ben AL, Cai XN, 2010. Analysis of synonymous codon usage in chloroplast genome of *Phalaenopsis aphrodite* subsp. formosana. Mol Plant Breed, 8:945-950.

[22] Yang GF, Su KL, Zhao YR, et al., 2015. Analysis of codon usage in the chloroplast genome of *Medicago truncatula*. Acta Prataculturae Sinica, 24:171-179.

[23] Zhang J, Chen GC, Xu JM, et al., 2016. Comprehensive selection for *Eucalyptus grandis* provenances and families. Journal of Tropical and Subtropical Botany, 24:280-286.

[24] Zhang ZY, Xiang DY, Xu JM, et al., 2016. Comprehensive analysis of growth, stem form and resistance to *Leptocybe invasa* of *Eucalyptus grandis* provenances. Forest Resources Management:107-111.

[25] Zhou M, Long W, Li X, 2008. Analysis of synonymous codon usage in chloroplast genome of *Populus alba*. J For Res, 19:293-297.

[26] Zhou ZP, Dang YK, Zhou M, et al., 2016. Codon usage is an important determinant of gene expression levels largely through its effects on transcription. Proc Nat Acad Sci USA, 26:e6117-e6125.

[27] Zhu SG, Xu CF, 2016. Biochemistry (4th Ed). Beijing: Higher Education Press.

Note: Figure translations are in progress. See original paper for figures.